The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# Attention-masking extended deep Q network（AME-DQN）reinforcement learning algorithm for combinatory optimization of smart-grid energy

Dinesh Bahadur Malla[*1,3] Tomoyuki Hioki[*2], Kei Takahashi[*2], Masaru Sogabe[*3], Katsuyoshi Sakamoto[*1,2], Koichi Yamaguchi[*1,2], Tomah Sogabe[*1,2,3]

[*1] Info-Powered Energy System Research Center,
[*2]Department of Engineering Science
The University of Electro-Communications, Chofu, Tokyo, 182-8585, Japan
[*3] Technology Solution Group, Grid Inc., Kita Aoyama, Minato-ku, Tokyo, 107-0061, Japan

Recently deep neural network-based reinforcement learning (DRL) methods, which demonstrated unprecedented success in game and robotic control, are gradually gaining attention to solve the combinatory optimization problem. However, effective operation in smart grid system has to be submitted to various constraints such as power demand-supply relation, lower and upper bound of battery electricity, market price etc. Because of these constraints, DRL algorithm is not efficient to get an optimized result. In this paper we address this issue by developing an attention-masking extended deep Q network（AME-DQN）reinforcement learning algorithm. Special focus was lied on the prediction ability of the trained AME-DQN model given various weather conditions and demand profile. These results were further compared with MILP results and finally we demonstrate that the AME-DQN are able to predict optimized actions which satisfy all the constraints while the MILP failed to meet the conditions in most of the cases.

## 1. Introduction

Defining the Energy system apart from the smart grid system is very difficult, improvement in research and artificial intelligence (AI) also makes system intelligent day by day. It's can be reasoning some authors even argue that it is "too hard" to define the smart concept [1]. The smart grid is an innovation that has the potential to revolutionize the transmission, distribution, and conservation of energy. Actually, the current electric power delivery system is almost entirely a mechanical system, with only limited use of sensors, minimal electronic communication and almost no electronic control [2]. Construction of efficient smart grid system is in principle a control optimization mathematical problem. Because of complexity wide range of methods have been proposed to tackle this challenge including linear and dynamic programming as well as heuristic methods such as PSO, GA, game or fuzzy theory and so on [3]. The mathematical process synthesis typically deals with the optimization with one objective and increase in parameter exponential increase in cost.

So, what is reinforcement learning? Reinforcement learning is a process where agents to learn optimal behavior under different conditions. Key concepts in reinforcement learning are state, action, reward, and policy [4]. The state refers to the state of the environment calculation at a given time. The action refers to the specific action taken by an agent, e.g. the direction and distance of an agent's movement within a given interval of time. The reward refers to the feedback signal (often a simple scalar value) given to an agent as a result of a specific action taken within a specific state. The policy links the states and actions of an agent and refers to the action(s) with the estimated highest reward value in any given state.

In this paper, we have an optimization objective for the energy grid system. Because of the complex system, we plan to solve small subsystem power cost optimization. Grid system has one household having Photovoltaic power production, having power storage battery and grid power supply. Where Production and consumption are not controllable but storing the power and use the storage can control, because of controllability optimization concept emerge. Small optimization for the household makes a huge quantity of the whole grid system, and subsystem optimization can connect with whole system optimization.

Contact: Tomah Sogabe, i-PERC, The University of Electro-communications, sogabe@uec.ac.jp

## 2. Model and Algorithm

Smart grid system based communication technology helps to know the current power demand, power production, battery SOC and other needed information. Basically digitalized electricity meter PV control system and sensing sensor plays a very important role collect the data and information. Base on the information we can optimize our system by taking what action makes out cost minimum. Our basic model is a small grid having one household for power consumption, one PV and one battery

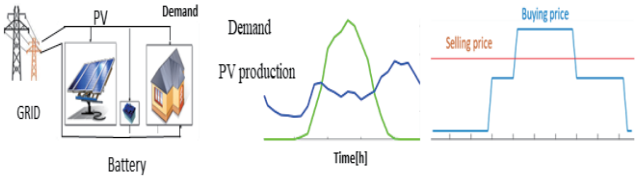


Fig.1 The basic model for power balance

for power storage and power supply for the consumption. The Q-attention-masking Algorithm is as below.

### 2.1 Reinforcement learning

We propose a framework that uses deep Q-learning to learn a high-level tactical decision-making policy, and also introduce, action-masking for the time of not constraints fulfill by the next-

state, a novel technique that forces the agent to explore and learn only a subspace of Q-values [7]. This subspace is directly governed by a constraints module that consists of prior knowledge about the system. Constraints of the problem and information from the same input state make the input data for the action-masking network. Not only does action-masking provide the tight integration between the two paradigms: learning high-level policy and using state action control, but also heavily simplifies the reward function and makes learning faster and data efficient. Not only Q-learning and action-masking are efficient to optimize our objective we use the learning from scratch, where agent updates their parameter by using search from scratch. We use epsilon-greedy for learning process which helps agent learn from scratch, the first agent selects action randomly and after the decrease in epsilon, agent use action by its learned result and agent able to optimize the goal.

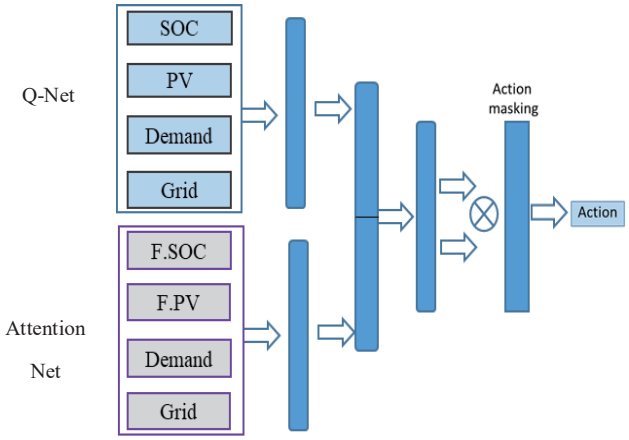


Fig.2 Attention-masking extended Deep Q algorithm

We use an attention network to parallel with normal Q-network, where attention network contains the information about the next state but not the state like next state. Attention network is a network, which informs the network about the current state and its impact to next state are fulfilling constrains or not. The cross-validation of state with attention state helps to make the good decision, and it helps the network to take the right decision based on the state to make next state. We start from scratch, with the help of the epsilon-greedy method we reduce the dependency of Q-network from random action. The decrease in epsilon helps our agent to calculate the action from the learned weight. Our attention states are so simple if constraint will satisfy in the next state they remain same, if not satisfy we just symbolized the state to -1.

After the layer of a network, we concatenate the state and future state result and calculate the action. For the purpose of hard constraints, we use an action masking process. Hard constraints are those constraints which are necessary to get the reward. After fulfilling the hard constraints, we have to fulfill the soft constraints which maximize or minimize the reward in the network and daily consumption cost of the electricity in the energy system. The network learns from energy optimization actions outcomes with the help of rewards by estimating the optimal Q-value function. Until the terminal time agent don't get a reward (like Monte Carlo samples), it gets a reward if satisfied all constraints if not it gets terminate punishment.

## 2.2 Mixed Integer Linear Programming (MILP)

MILP is a mathematical optimization program in which some or all of the variables are restricted to be integers. The mathematic optimization methods can divide into three parts where the first is a single numerical quantity objective function which is to maximized or minimized. The second is a collection of variables which are quantities whose values can manipulate in order to optimize the objective. The third is a set of constraints which are restrictions on the values that the variable can take. Here in work our objective function is as below: $minimize\ \sum_{h=1}^{48} C_{buy}P_{buy}(h) - C_{sell}P_{sell}(h)$ and constraints are :

$$P_{demand} = P_{buy} - P_{sell} + P_{battD} - P_{battC} + P_{PV} \ldots (1)$$

$$P_{battE}(1) = P_{battE}(48) = P_{battEmax} \qquad \ldots (2)$$

$$P_{battE} + dt\ P_{battC} \leq P_{battCAP} \qquad \ldots (3)$$

The Constraints 1 means power demand needs to be equal with power buy minus power sell and adding PV with a battery charge. The Constraint 2 for battery state of charge (SOC) at end of the day is equal to the start of the day. The Constraints 3 is for battery capacity is always greater than battery current adding with powered charged in the battery.

## 3. Results and discussion

We already discuss RL algorithm in section 2.1, Base on that algorithm we obtain the different result that we discuss in this section. First, shortly describe coming for this algorithm. The field of optimization is totally obtained by the mathematical optimization methods, which is very good for one objective optimization. But it is not cost effective and not good enough after the increase in a parameter and objectives. So, the optional method is reinforcement learning, but RL is very weak in constraint fulfillment but very cost effective. We purpose this reinforcement algorithm for constraints based optimization problem. The obtained results are here below:

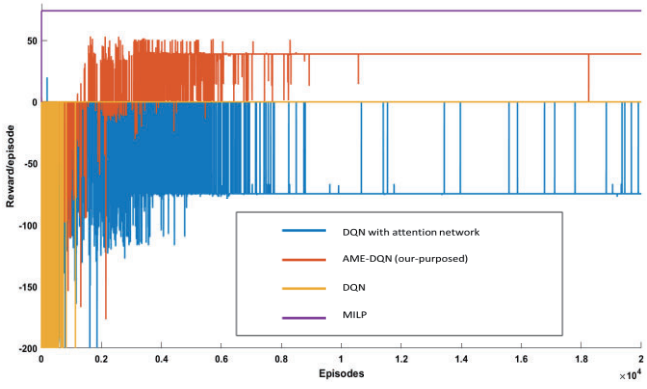


Fig.3 Reward per episodes of different R L algorithm

The Fig.3. is the reward results per episodes of our suggested algorithms. Here we used a reinforcement learning algorithm to optimize the power cost per day. In reward plot 0 mean some or one constraint is not fulfilling. The negative reward means a daily cost to pay for electricity consumption and positive means income from the electricity transaction. The DQN is the similar algorithm presented by the deep mind team[5]. The DQN Attention Mask network is the algorithm combination of DQN, attention, and masking, where attention network used in language processing techniques and masking is also one of the actions clarifying technique used in RL recent year. The DQN Attention network is the combination of DQN and attention network. The above graph clearly presents that, only DQN algorithm is unable to satisfy the constraints. In the beginning time a random process of taking action helps to satisfy the constraints but decrease randomness, it does not satisfy the total constraints. The other DQN with Attention is able to satisfy the constraints but it is also unable to minimize the cost and DQN with Attention and mask network is quite satisfactory to compare with DQN and DQN with Attention network.

## 3.1 AME-DQN test

Reuse is the main benefit of RL, so it is cost effective, and it is efficient to optimize the similar nature problems with the previous learned network or agent. For the test purpose, we defined the similar nature power problems and use the above discussed AME-DQN agent to solve the problem, the results are discussed in this section. In Fig.4. PV production curve represents the low production of PV or cloudy weather PV

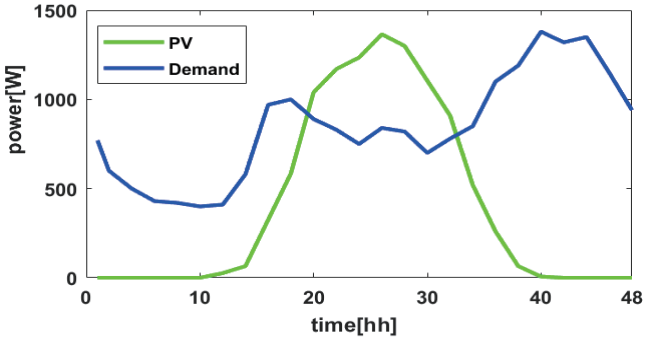


Fig.4 Demand and PV production curve

production. The total production of the day is 19959[W], and demand is 40400[W]. The optimization result of MILP and test result of RL plotted in the below Fig.4, Fig.5. The MILP mathematic optimization algorithm [6], and the RL side only use the learned weight to solve the problem. Here in this test, only PV production is different from the above-solved problem, but MILP needs to recalculate the problem to solve or optimize the result but not to the RL. MLP optimized result is lower than the RL test optimized cost for the day, but this is the result of weight used where MILP optimized result is 74 JPY/day, and RL has only 38JPY/day. At the test time, all the constraint is fulfilled and it also near to optimized result also.

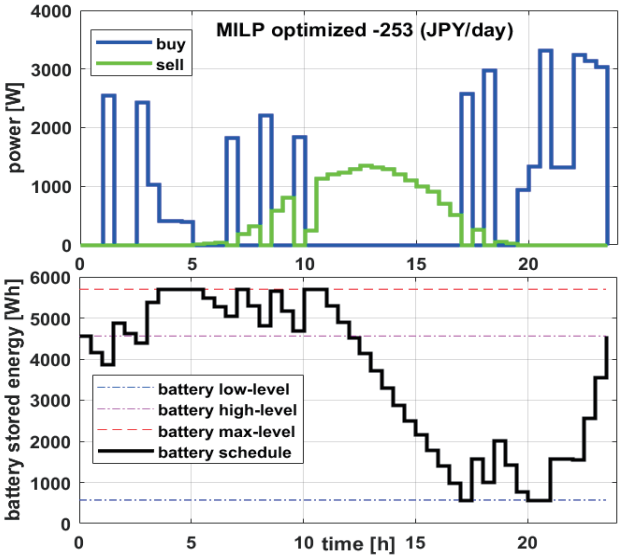


Fig.5 Buy and sell schedule from MILP optimization

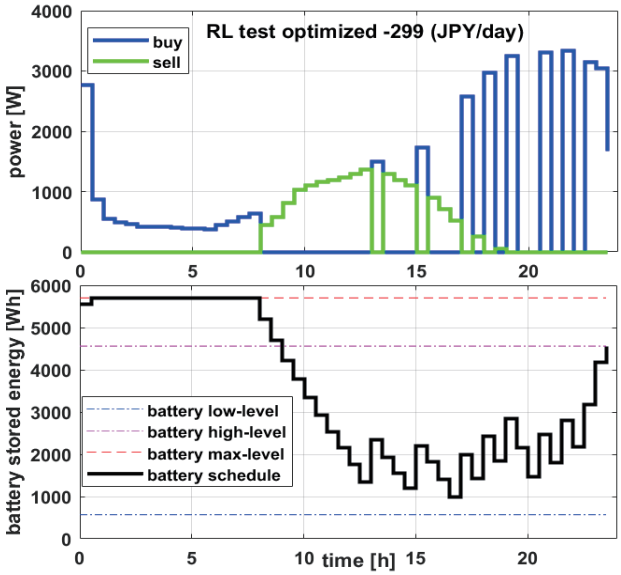


Fig.6 Buy and sell schedule from RL test

## 3.2 AME-DQN and MILP test

In this section, we present the same power system learned AME-DQN and MILP. MILP test is a quite unfamiliar term, we used MILP optimization time actions for the same time step of the problem to test, and AME-DQN is using the learned weight to forecast the test. Here in this work, we used the last optimization time actions of MILP because all PV production and demand are not the same. Both of the algorithms have learned the same problem whose total reward is plotted in Fig.3. This time test problem demand and PV production curve is in as Fig.7. MILP and AME-DQN results are plotted in Fig.8 and Fig 9 respectively. The problem contains different PV production curve and higher demand of the 24 hours.

Form the test results we can see a different scenario, which is not an unbelievable pattern from the MILP optimization. But in the test time we haven't calculated the optimized result, we only

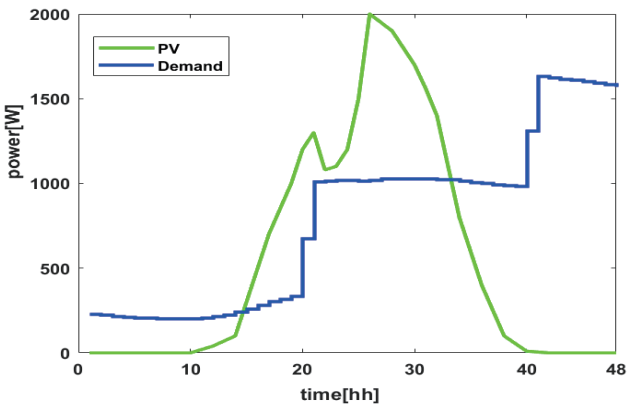


Fig.7 Demand and PV production curve

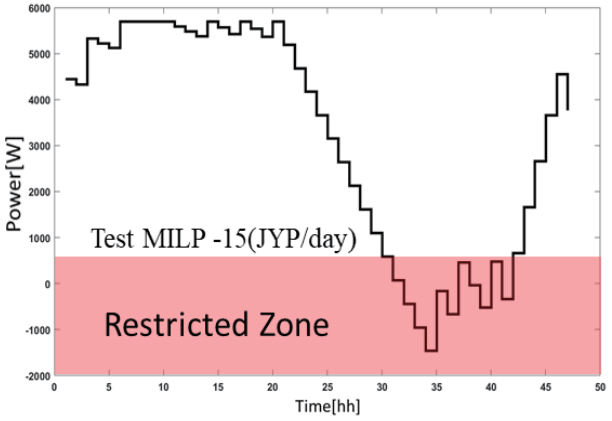


Fig.8. test time battery schedule by MILP

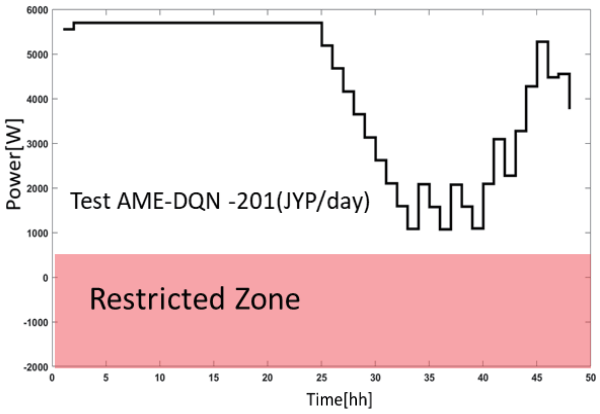


Fig.9. test time battery schedule by AME-DQN

predict on the base of optimization time scenario. So, MILP is not for prediction purpose, but average from predicted different scenario can make it better. The test result shows that MILP is not obeying the lower bound constraints at a test, but AME-DQN fulfills all the constraints.

## 4. Conclusions

In this paper, we present a reinforcement learning method for smart grid optimization. From the different result discussed above in results and analysis section, the agent was able to catch the feature involved in the balance of load demand, PV power surplus and battery discharge/charge, as well as grid, integrate. The agent successfully learned how to tune its action profile to maximize the reward function during training. The RL agent satisfies all the constraints, which is one step toward the optimization. Our learning agent is not able to get the global goal but can be useful for similar kind of problems. So, there are many beneficial parts for using the RL agent than the using MILP or another mathematical optimization result. Mathematical optimizations are good for one objective optimization but increasing in data, parameter and objectives make it incompatible. If we want a multi-objective result that is only fulfilled by a reinforcement learning agent, it is also cost-effective and we can model without knowing the problem from the root. The RL for an optimized result is difficult but not impossible. RL agent can optimize the problem from scratch now our RL agent fulfill the constraints which are difficult to fulfill so far. The current work can continue by uniting more power sources in the future. Also, works focus on global optimization and optional roots search for optimization. This research helps us to know about the constraint application in RL learning process, how to define soft and hard constraint as well.

## References

[1] Miller, J., “The Smart Grid – How Do We Get There?”, Smart Grid News, June 26, 2008

[2] http://smartgrid.epri.com

[3] M. R. Alam, M. St-Hilaire, and T. Kunz, "Computational methods for residential energy cost optimization in smart grids: A survey," *ACM Comput. Surv.*, vol. 49, pp. 22-34, Apr. 2016.

[4] L.A. Bollinger and R. Evins. “Multi-agent reinforcement learning for optimizing technology deployment
in distributed multi-energy systems”. EMPA20160705

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. 2015. “Human-level control through deep reinforcement learning.” Nature 518(7540):529–533.

[6] M.J. Vahid-Pakdel, Sayyad Nojavan, B. Mohammadi-ivatloo, Kazem Zare. Stochastic optimization of energy hub operation with consideration of thermal energy market and demand response. Energy Conversion and Management 145 (2017) 117–1

[7] Mustafa Mukadam, Akansel Cosgun, Alireza Nakhaei, Kikuo Fujimura. “Tactical Decision Making for Lane Changing with Deep Reinforcement Learning”. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.